DOCUMENT RESUME

ED 295 991 TM 011 843

AUTHOR Coffman, William E.
TITLE An Historical Perspective on Key Measurement Issues: Individual Differences, Errors of Measurement, and Accountability.
PUB DATE Apr 88
NOTE 27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Accountability; Educational Assessment; *Educational History; Educational Trends; Elementary Secondary Education; *Error of Measurement; *Individual Differences; Literature Reviews; *Program Evaluation; Scores; *Test Interpretation; Test Use

ABSTRACT

Given the wide individual differences among any group of students and since measurements are always accompanied by errors, the question of how tests should be used in assessing the quality of an educational program is considered. The ways in which educators have dealt with this problem are reviewed, from the systematic examinations of the question beginning in the 1930s through the grasp of the problem today. A better understanding of the evaluation process as proposed in 1951 by W. Cook is presented. Since that time educational measurement has been characterized by alternating emphasis on the development of excellence and a focus on minimum essentials. There currently is no general agreement on reference standards, and no simple answer to the meaning behind test results has been found. Five exhibits present sample data illustrating the effects of individual differences among students' test scores. A personal perspective is provided throughout the discussion. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295991

SYMPOSIUM

AN HISTORICAL PERSPECTIVE ON KEY MEASUREMENT ISSUES

Individual Differences, Errors of Measurement, and Accountability

William E. Coffman
University of Iowa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

William E. Coffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the
National Council on Measurement in Education
April 6, 1988

TM011843

AN HISTORICAL PERSPECTIVE ON KEY MEASUREMENT ISSUES

Inidividual Differences, Errors of Measurement, and Accountability

William E. Coffman
University of Iowa

In his letter acknowledging my acceptance of his invitation to participate in this symposium, Neal Kingston indicated that each of us should "choose an issue with which he has been concerned throughout a major part of his career...(and)...discuss the changes he has seen in our understanding of that issue and the way in which we have dealt with the issue." This is the issue I have chosen: Given that there are wide individual differences among any group of students and that measurements are always accompanied by errors, both random and systematic, how should one go about using tests in assessing the quality of an educational program?

In reviewing the issue, one might, of course, begin by carrying out a systematic review of the literature. But this is not only a massive undertaking; it probably wouldn't provide a very accurate picture of "the way in which we have dealt with the issue." As Jaeger (1987) reminded us in his NCME presidential address last year, we often do not appreciate the significance of frontier ideas until years after they appear in the literature. Scates (1938) put it even more strongly fifty years ago:

> Each generation seems to discover for itself teleological and methodological concepts which it brands as new, or progressive, even though these very ideas may have been formulated and voiced centuries or millenniums earlier. It is difficult to know what is new; most ideas are new only to individuals. It appears, however, that there

---

Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, April 1988.

(Coffman, 1988)

> are strong movements in education today which are actually affecting practice in conventional schools in ways which heretofore were only talked about....(p. 523)

Looking at what is happening today, one may well ask whether Scate's conclusion would still apply or whether the strong movements of today are away from directions encouraged by research findings. It is true that slowly and gradually, sound ideas do influence practice, but the path of progress is a crooked one with regressions as well as advances.

For the present generation of the educational measurement family, then, it might be better to describe my own pilgrimage on the road to an understanding of the issue, with references here and there to contributions the literature was making - or might have made if I had known about them.

The year I made the decision to enter the field of education, 1931, was the first year of publication of the Review of Educational Research; and two years later the February issue carried the first review on the topic "Educational Tests and Their Uses" (Wood, 1933), a review that cited 467 references. The Education Index first appeared in 1929, and the first bound volume in the University of Iowa library (January 1929 - June 1932) contains entries reflecting the fact that there was already a history of educational measurement (Fish, 1930; Malin, 1930).

The periodic reviews, however, also document the fact that there is a lag between the first appearance of an idea and its general dissemination. The heading "Evaluation" first appeared in the second bound volume of the Education Index (July 1932-June 1935), but there was only one reference, and

(Coffman, 1988)

the heading "Evaluation" did not appear in the Encyclopedia of Educational Research until the revised edition (Monroe) in 1950.

The span of my own professional career covers the period since these systematic reviews first appeared. The first third of the period I was a student, classroom teacher, or principal, a user of tests rather than a measurement specialist. Fortunately, my home state of West Virginia, where I expected to be teaching, required a course in tests and measurements for certification as a secondary school teacher. Since I was enrolled in a college in Ohio, and since Ohio did not require such a course, I found myself completing the requirement through individual study. Thus, when I began teaching, I knew that test scores were subject to errors of measurement, that there was experimental procedures for estimating the magnitude of those errors, and that there were certain standardized tests with norms that might provide a frame of reference for interpreting results.

The required course in educational psychology had also introduced me to the concept of individual differences, but like many of the concepts encountered in textbooks and lectures only, the course did not prepare me for the range of differences I encountered once I began teaching: one student who never scored other than perfect on any examination throughout the year and another completely unable to read or write and the others scattered between with a heavier concentration in the middle. And this wasn't because I was making tests designed to produce a normal distribution. The professor who taught general methods might well have called the course "How to Succeed in Schools as They Are," and so I found myself including in the examinations both objective and essay questions all of which I thought asked about what had been included in the instruction.

(Coffman, 1988)

The term "accountability" does not appear in the literature of that time, but the idea that teachers and administrators were responsible for reporting to the public, including parents and school boards, how things were going in the school was certainly part of the educational climate, as was the idea that one ought to be checking up on one's self. Examinations could provide both a basis for reporting to parents and for letting the teacher know whether or not what was being taught was also being learned. The results were often disappointing.

By 1941, I was a principal of a small consolidated elementary school, and I had managed to find enough money to purchase a battery of standardized tests for the school. (The per pupil expenditure in the district school budget was less than $100 per pupil per year, so such luxuries as tests or library books had to come from funds raised locally.) Exhibit I is a summary table I prepared for the testing program in the spring of 1941. From today's perspective it certainly is primitive; but it does indicate some awareness of the problem of errors of measurement (we were looking at averages, not at individual scores); of individual differences (in a school where score distributions were positively skewed, it seemed preferable to look at medians rather than means, although means would have given higher averages); and of the need to make adjustments for differences between the general abilities of these pupils and those of the norms sample (socioeconomic level of the school community was clearly below average, and the typical pupil entering the first grade was obviously deficient in language development.)

We didn't know at the time, but we were recognizing the need for a "headstart" program even if it had to start in the first grade. Our solution

was a three year primary program with pupils moving through a sequence on a continuous progress pattern with some pupils spending as much as half a year in prereading activities and some needing a fourth year to complete the sequence. In comparison with the performance on the general ability test (I believe it was the Otis) the performance on tests of subjects taught in school were interpreted as indicating that the primary program was on the right track. It seemed clear, on the other hand, that the program in grades 5 through 8 left something to be desired.

Before we could solve the problem of what to do about the program in the upper grades, or to collect longitudinal data to see if the succcess of the primary program would carry over into the upper grades, I found myself in the army teaching English to Puerto Rican recruits. This was my first experience with a minimum essentials program; the task was to teach enough English so that recruits could function in units commanded by English speaking officers. The problem of individual differences was tackled by dividing recruits into ten instructional levels on the basis of pretest scores. In theory the top groups should complete the program relatively quickly while lower groups would require a longer period of instruction. Class size was 12. Instruction proceeded in one-hour sessions alternating with one-hour periods of basic field instruction. Each instructor had only a single class. Thus, in an eight hour work day, the instructor had four hours for preparation. One might well conclude that if educational budgets were as supportive as the military budgets, we might indeed solve the problem of teaching minimum essentials.

(Coffman, 1988)

The only fly in the ointment was that so long as recruits did not complete the program, they would not be sent into combat units. The concept motivation began to take on new meaning for me.

Following the war, I returned to a systematic study of education and found much in the literature that contributed to a clearer understanding of the issues I have been struggling with on the firing line. I was now planning to teach at the college level and found that here too I would have to deal with the problem of individual differences (Learned & Wood, 1938). I found that Lindquist (1951) had been concerned with the relationship between measurement and instruction at the public school level and that Tyler (1932) had been doing the same at the college level and that both were kindred spirits.

Walter Cook (1951) stated much better than I could what I had come to believe regarding the function of testing in the schools:

> As long as all the children of all the people remain in school, it will be impossible to reduce the variability of instructional groups significantly through promotion policies. If rigorous policies of promotion are adhered to, the efficiency of the school is reduced through the accumulation of low-ability pupils and the lessening of educational opportunities for the more able.(p. 21)
>
> .....
>
> ...if the responses to be learned are sufficiently simple and the goals that have been set so limited that a high proportion of the group can master them during the period of learning, the variability of the group becomes less; if the task is complex and the goals unlimited, so that the abilities of the most apt members of the group are taxed during the period of learning, the variability of the group increases.(pp. 21-22)
>
> .....

(Coffman, 1988)

> The most serious results of emphasizing limited goals in education is that what is learned is frequently of little value and is retained for but a short period.(p. 22)
>
> .....
>
> Unless learning involves differentiation and integration of old and new responses into a problem-solving type of mental process or into an organized behavior pattern, it has little permanence or value. How was it learned? is the important question.(p. 23)
>
> .....
>
> The permanent results of the educational process are measured by such tests as: (1) vocabulary, (2) reading comprehension in the natural sciences, social sciences, and literature, (3) problem-solving in mathematics and the sciences, (4) ability to use the library and basic reference materials, and (5) ability to write and speak effectively. It is with reference to such objectives that great heterogeniety in achievement exists, and the more effective the instruction, the greater the heterogeniety.(p. 23)

Cook's analysis points up a key issue in the design of an evaluation program; what one includes in the design and what one expects as outcome depends on one's view of what schools should be aiming for. If the goal is to insure that some limited set of concepts and skills will be learned, at least long enough to pass that test, the measurement will focus on those limited essentials and the effective school is one that has an increasing percentage of pupils over some selected threshold. On the other hand, if the goal is to maximize the learning of all pupils with the focus on abilities for dealing with an ever-increasingly complex culture, then evaluation will have these characteristics: (1) Measurement will focus on the permanent outcomes of the educational process in terms of complex skills and understandings (although for what we now think of as formative evaluation, diagnostic tests have a significant

(Coffman, 1988)

place.) (2) Evaluation should tell us what is happening over time; a single assessment is never enough to show what is happening. (3) It should reflect what is happening to all pupils; measures of central tendency alone will not do the job.

The history of educational measurement since the publication of Cook's analysis has been characterized by alternating emphases on these two views of goals. The launching of sputnik at the end of the 1950's led to a renewed focus on the development of excellence in the 1960's. Evidence that some pupils were not learning to read or to calculate, particularly in inner city schools, led to a focus on minimum essentials in the 1970's. Today there is again concern about the development of thinking skills.

Sometimes the concern seems to focus on both at the same time but without the comprehensive view of the teaching process expounded by Cook. People seeking a simple solution to the complex problems of teaching and learning propose to raise standards by promoting only those students meeting some set standard and at the same time propose to eliminate high school dropout by the year 2000. They haven't read Cook.

To read some of the articles appearing in educational journals today, one might conclude that individual differences are of little significance in the issue of school quality; all one has to do is to hold the professionals running the schools responsible for the failure of pupils to meet standards and results will improve. It is true that some, at least, of the average differences among pupil groups result from differences in opportunity and may be eliminated by fairer treatment; but there are also individual differences within whatever groups are identified, differences arising from complex

(Coffman, 1988)

multieffects still not well understood and not likely to be overcome in the foreseeable future. As a recent summary of the topic (Peterson, 1982) indicates, individual differences are still with us:

> Students differ. Within a given fifth-grade classroom, for example, students may differ in ability, motivation, attitude, personality, learning style, sex, ethnicity, and socioeconomic status. Moreover, classrooms appear to be becoming more diverse rather than less diverse. The passage of Public Law 94-142, the increased attention to students with limited English-speaking ability, and the federal and state emphasis on ensuring equal treatment of students have resulted in a situation in which teachers confront students with a wide range of characteristics within their classrooms.(p. 844)

The current literature also documents increasing emphasis on teaching students how to think more effectively. Some of the articles seem to imply that the possibility of teaching thinking skills is a new idea. Obviously, the writers haven't read Cook, or a host of other educators over the years (Coffman, 1987). Let us hope that efforts to improve education by accountability mandates do not lead to a swing back of the pendulum to concentration on minimum essentials. As one article described the emphasis of the previous swing (Kirst & Wirt, 1982):

> Between 1966 and 1976, thirty-five states passed such laws, with fourteen adopting a comprehensive system. Some 4,000 pieces of writing about it were published. Laws dealt with budget format, teacher evaluation, curriculum assessment to meet standards, the setting of LEA objectives and minimum competency standards for high school graduation. As one observer of this development illustrates from Florida,

(Coffman, 1988)

> "State assessment tests in certain subjects are publicized through school-based parent councils. Statewide tests are also used for high school graduation. Students must score in the 70th percentile statewide to obtain a diploma. Student test scores are related to teacher evaluations. School districts fear the adverse publicity that can come with publication of test scores lower than other districts." (p. 1774)

The measurement literature beginning in the middle 1950's reflect a growing concern about the possibility that the results of tests used for accountability might be misinterpreted, particularly as a result of errors of measurement or of oversimplified analyses of the data (Lord, 1956; McNemar, 1958; Harris, 1963; Lord, 1967; Werts & Linn, 1969; Dyer & Patton, 1969; Cronbach & Furby, 1970; Werts & Linn, 1971). These developments provided the measurement specialists with ways of avoiding misinterpretations. They were, however, of little help in communicating with the naive general public, or even with the typical teacher or administrator. The problem is that individuals operating from common sense suspect that the statistics used to adjust tests scores to make valid comparisons are actually designed to make the schools look better than they are and busy teachers and administrators have no time to master the complicated statistics.

A number of studies conducted by some of my students or by me were aimed at the development of simpler approaches that might incorporate the results of the theoretical developments while still being understandable to the public. One of these studies (Coffman, 1960) arose as a result of criticisms of the use of test score averages in evaluation studies. The claim was that growth patterns for individual students were idiosyncratic, that is, that by averaging scores one obscured what was happening to individual

(Coffman, 1988)

pupils (Ketchan & Lafitte, 1959). The only legitimate function of tests, according to these critics, was to provide teachers with information about the status of students at the time of testing so that instruction could be adjusted to individual differences.

I supposed that many of the so-called unique growth patterns represented primarily errors of measurement, and if these could be discounted in some way, growth patterns would appear less idiosyncratic.

The results of the study, conducted using battery mean scores of two groups of pupils who had been tested near the end of four successive school years on the Stanford Achievement Test Partial Battery, are pictured in Exhibit II. Pupils were divided into upper, middle, and lower thirds, not on the basis of initial scores but rather on the basis of the sums of the four measures. In summarizing the results, I wrote:

> When longitudinal data are accumulated and combined to minimize the effects of errors of measurement, a reasonable regularity appears. There is no support in these data for the proposition that growth patterns are highly individual and can provide no basis for predicting future performance. On the other hand, there is every indication that decisions based on a single test score, even a highly reliable battery median, should be tentative, that with each new measurement over time, the picture becomes clearer. The individual lines are characterized by the back and forth fluctuations of error, not by continuously accelerating or continuously decelerating curves of changing growth rates. With each new measurement, the underlying pattern becomes easier to approximate. The growth lines for the top, middle and bottom thirds indicate the divergence we have come to expect under learning conditions which recognize the presence of individual differences and which stimulate each child to explore

(Coffman, 1988)

> in meaningful ways the interesting world around him and the interesting world of books.(pp. 145-146)

From this study I concluded that by looking backward from an accumulation of several years of measurement one could obtain a reasonably good picture of what was happening to pupils in a school without using elaborate statistical procedures. However, looking at only one or two measurements presented a somewhat different problem. On the appearance of Lord's paper on the measurement of growth (1956) I wondered what the implications might be if his equations were applied to data I had been working with. My analysis is pictured in Exhibit III (Harris, 1963, p. 29).

As you can see, the analysis does not change the location of the points in the scatter plot but simply changes the location and slope of the reference lines. Also it changes the conclusions we might draw regarding the relative gains of pupils at different parts of the socre scale. While the original data suggest that both good students and poor students are gaining about the same amount in terms of score scale, the application of Lord's equations indicates that good students are learning relatively more than the poor students, a conclusion already demonstrated in Exhibit II. Thus, the teacher or principal, using the procedures of Exhibit II will be less likely to reach incorrect conclusions than one using the procedure of Exhibit III without the adjustments in the reference lines.

One question that needs to be examined before placing too much emphasis on changes in test scores for a school from one time to another, regardless of adjustments, is whether or not results are consistent from cohort to cohort. If, for example, we find a particular pattern between Grade 5 and Grade 7 for pupils in Grade 7 this year, can we expect the same pattern for next

(Coffman, 1988)

year's seventh grade pupils? This was the question investigated by Chen (1974) in her master's thesis at Iowa.

Using test scores on the five subtests of the Iowa Tests of Basic Skills two years apart (Grades 4 and 6 in four schools and Grades 5 and 7 in twenty-one other schools), Chen applied analysis of covariance to the data from successive cohorts in the 25 schools. Not only were the order of joint means different from cohort to cohort, but there were differences also for the successive slopes for certain of the measures. She concluded her report with these words:

> For persons who are interested in carrying out school accountability studies using standardized achievement tests, stability of the test results over years should be examined before the conclusion is drawn. To base a conclusion on a single year's data is likely to be unfair. First, there are errors of measurement, and usually the sample available for a single cohort is small. Also, school treatments, cohort differences, community policy, community population mobility, and school staff mobility may all contribute to explaining the results. Rather than use the test results to classify schools as effective or ineffective it might be better to use them as feedback for guiding school policy; that is, to help the school decide where to put instructional emphasis on their next year's work and thus change the results for next year. In that case, the test results would no longer be functioning as a basis for praising or blaming schools, but rather to guide professional action.(pp. 54-55)

A skeptical reviewer of Chen's work might say it's not surprising that a student at Iowa, using data from the Iowa Testing Program would come to the same conclusion regarding the proper use of tests in 1974 as did

(Coffman, 1988)

E.F. Lindquist in 1936 when the program was initiated. But her data are not so easily dismissed. Given that schools differ in a variety of ways and that there are errors of measurement, no statistical analysis can take the place of professional judgments in reaching valid conclusions from evaluation data (Wests & Linn, 1969). Unfortunately, few of the legislators who propose accountability legislation, or small town doctors with time to collect test reports from all fifty states, or journalists who prepare reports of test results for newspapers, are qualified to make such judgments. Watch out! The next thing you know they'll be wanting to publish the average tests results from the files of a sample of physicians in each state to let us know which of the states have the highest quality medicine.

The studies I have discussed to this point have been concerned with how one may look at test results from a particular school. But what about the question of standards? In one sense, since each school is unique, one might argue that the most important question is whether or not results are improving over a period of years, and it is only the data from the single school that are needed to answer such a question. On the other hand, one does wonder whether results in other systems might be better or worse in some sense. Without a reference standard, one's search for a conclusion is clearly limited.

A recent report of ERIC Clearinghouse on Tests, Measurements, and Evaluation (Frederick, 1987) indicates that there is no general agreement on this issue. Different evaluation studies use different standards, and the methods of analyzing the data vary depending on the criteria adopted. The report does make a recommendation that isn't too different from the approach I used in the 1960 study: Divide the group into three levels, obtain data

(Coffman, 1988)

over a number of years for each cohort, and compare the average scores from year to year with the corresponding percentile ranks for the norms of the test. However, the recommendation offer little advice about how one might decide whether the results are favorable or unfavorable, given the fact that school populations differ from each other in various ways.

The issue of standards was tackled in a recent doctoral study at The University of Iowa by Olsen (1982). He reasoned that one ought to have the flexibility to consider various "standard" groups to which results for a school might be referred, for example, the national norms group, the state norms group, the local system norms group, or even, perhaps, a group selected because it was judged to be an acceptable standard.

Given measures at two points in time, the individuals in the reference group and in the group being evaluated might be pictured by dots and pluses as in Exhibit IV. Olsen argued that one might summarize such a result by plotting lines joining equal percentile points in the joint distributions as illustrated in Exhibit V. However, the task of judging the significance of differences between the two lines would be next to impossible; a more simplified summary was needed.

Olsen's review of the literature concluded that there were several lines of theoretical developments all of which concluded that the appropriate line for summarizing a joint bivariate distribution based on fallible measures was a line through the joint mean with a slope equal to the ratio of the posttest standard deviation to the pretest standard deviation. Given the means and standard deviations for the local group and those for the comparison group, an evaluator could very quickly prepare a chart showing the relative position of the two lines.

(Coffman, 1988)

Since pupils in a particular school could in no way be considered a representative sample from the standard population, there was the problem of how to decide whether or not a line for a local school was different enough to be considered significantly better or worse than the standard, either with respect to the joint mean or to the slope of the line.

Suppose, however, Olsen reasoned, schools in the reference population had equally effective educational programs and differed only in the proportion of pupils falling into each of the quarters of the joint standard population, an assumption that would lead to distributions of test scores within schools very much like those observed for collections of real schools. One could then carry out a simulation study by drawing such samples and preparing a sampling distribution of school lines and from such a distribution obtain estimates of the sampling error of slopes and of deviations of joint means from the reference line.

The simulation study was carried out using data from four successive years for a large city school system. The distributions for the simulation were indeed less variable than those for the actual schools in the system, indicating that schools actually do differ more than one might expect if educational influences were actually the same from school to school in the system.

More recently Olsen's study has been replicated using data from a large number of schools in the Iowa Basic Skills Testing Program (3 cohorts, 3 testing periods, 4 succeeding testing years). The estimates of error are slightly smaller than those found by Olsen, but otherwise, the results are comparable.

(Coffman, 1988)

Once differences are identified, however, one still faces the task of deciding what factors are responsible for the differences that are observed, and here professional judgments are involved. Unfortunately, there is no simple answer to the meaning behind test results.

EXHIBIT I

(Prepared by W.E.Coffman, May 1941)

Median Scores of Dixie Pupils on Metropolitan Achievement Tests - 1941

↓ Test Grade →	1	2	3	4	5	6	7	8	Med. Mos. diff. from Norm.
I.Q.	84.5	91	89	95	98	84	84	93	Using National Norms
M.A.	1° 6-0 -8	(2') 7-3 -7	(3°) 8-2 -8	(4') 9-5 -7	(5') 11-3 -1	(5') 11-3 -11	(5') 11-5 -11	(7') 13-4 -16	-9
Ave. Achievement	B^3 -15	2^7 -1	3^7 -1	4^9 +1	5^1 -7	5^7 -11	6^7 -11	7^2 -16	-4
Reading	B -18	2^7 -1	3^8 10	4^8 10	5^0 -8	5^8 -10	6^0 -12	7^1 -17	-5
Vocabulary	1^9 -15	2^8 10	3^8 10	5^2 +4	5^4 -4	5^7 -11	6^3 -15	6^8 -20	-4
word recognition	2^7 -19								
Ave. Reading	2^4 -16	2^7 -1	3^8 10	5^0 +2	5^1 -7	5^7 -11	6^2 -16	7^1 -17	-4.5
Arith. Fund.		2^8 10	3^6 -2	4^5 -3	4^8 -10	5^8 -10	6^6 -12	6^8 -20	-8
Arith. Prob.		2^7 -1	3^3 -5	4^2 -6	4^8 -10	5^5 -13	6^6 -12	7^0 -18	-9
Ave. Arith.	2^3 -15	2^8 10	3^5 -3	4^4 -4	4^7 -11	5^8 -10	6^6 -12	6^8 -20	-7
English			3^7 -1	5^4 +6	5^9 +1	6^1 -7	7^6 -2	7^7 -11	-2
Literature				5^1 +3	5^9 +1	6^4 -4	6^7 -11	7^1 -17	-6
History				4^9 +1	5^3 -5	6^3 -5	7^4 -4	7^5 -13	-5
Geography				4^6 -2	5^1 -7	5^5 -13	7^2 -6	7^7 -11	-8
Spelling		2^4 -4	4^1 +3	4^7 -1	5^5 -3	5^8 -10	5^9 -19	6^5 -23	-8
Achievement - M.A.	+13	+6	+7	+8	-6	0	+8	-6	+3¾

BEST COPY AVAILABLE

EXHIBIT II

(Coffman, Seventeenth Yearbook, NCMUE, 148-151)

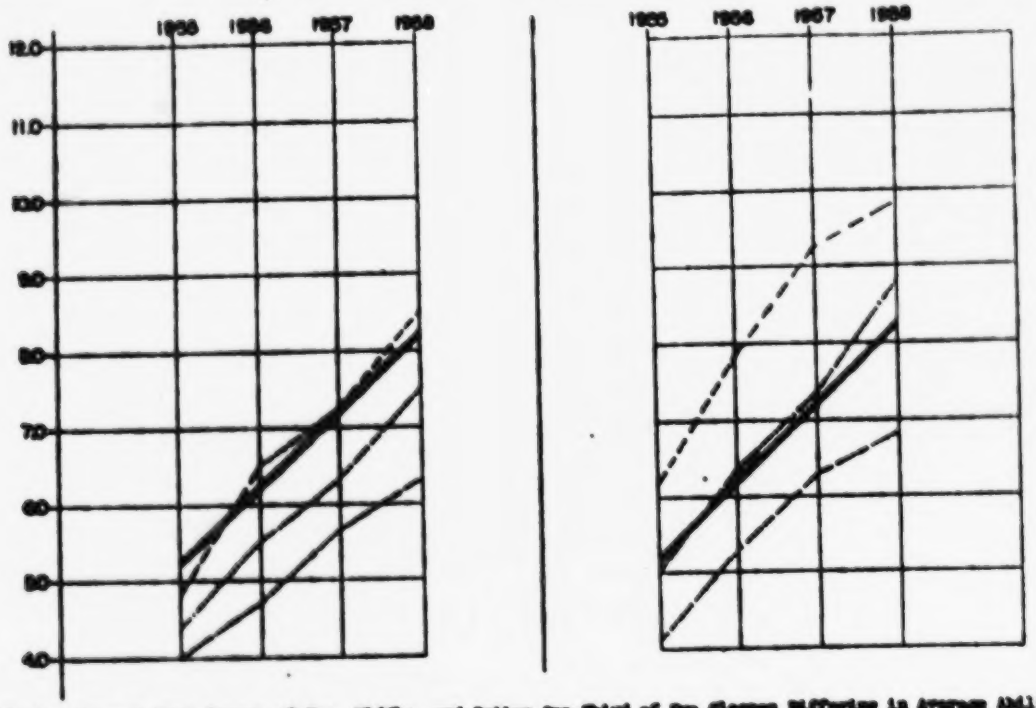


Fig. 1. Average Test Scores of Top, Middle, and Bottom One-Third of Two Classes Differing in Average Ability

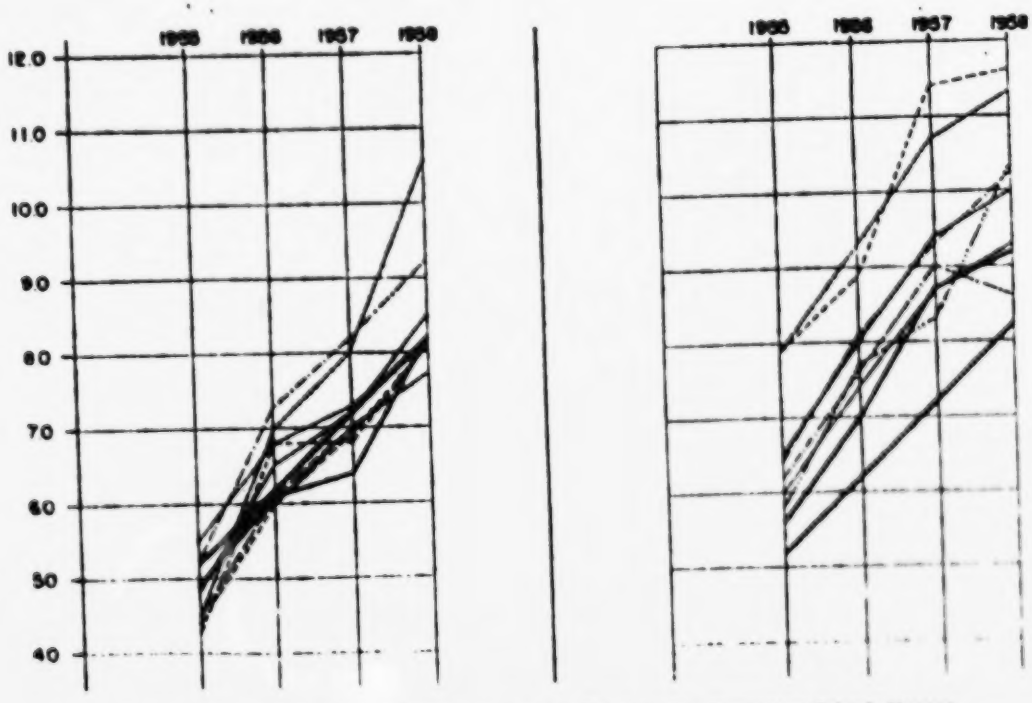


Fig. 2. Test Scores of Individuals in Top One-Third of Two Elementary School Classes

EXHITIB II (cont.)

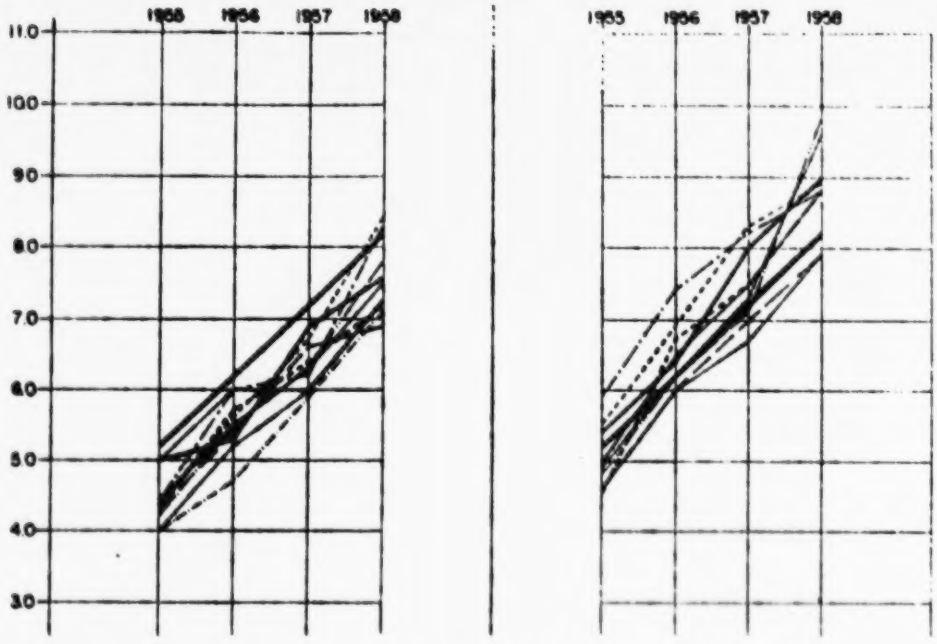


Fig. 3. Test Scores of Individuals in Middle One-Third of Two Elementary School Classes

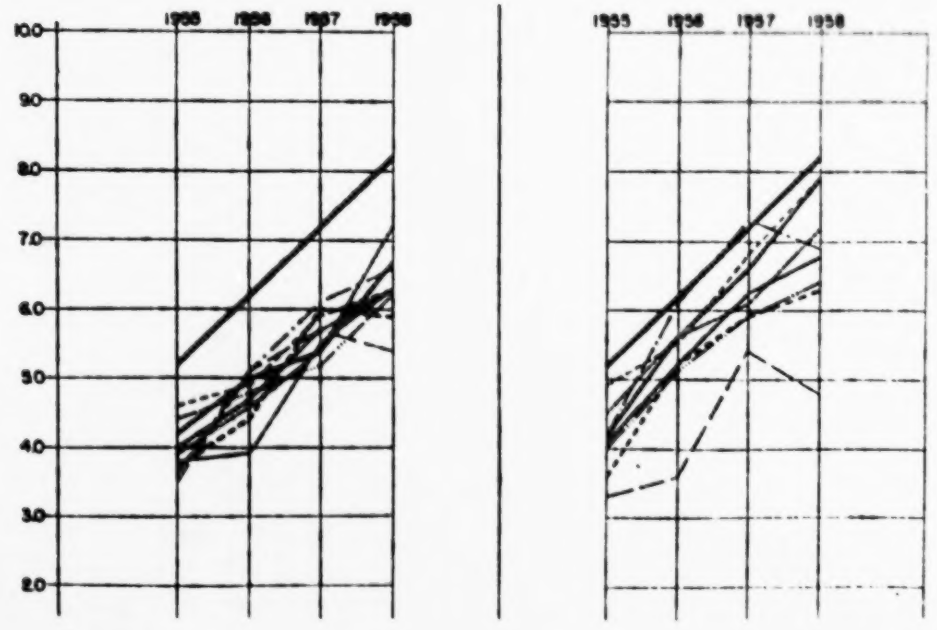


Fig. 4. Test Scores of Individuals in Bottom One-Third of Two Elementary School Classes

EXHIBIT III

(Lord, In Harris, C.W. (Ed.).Problems in Measuring Change, p.29)

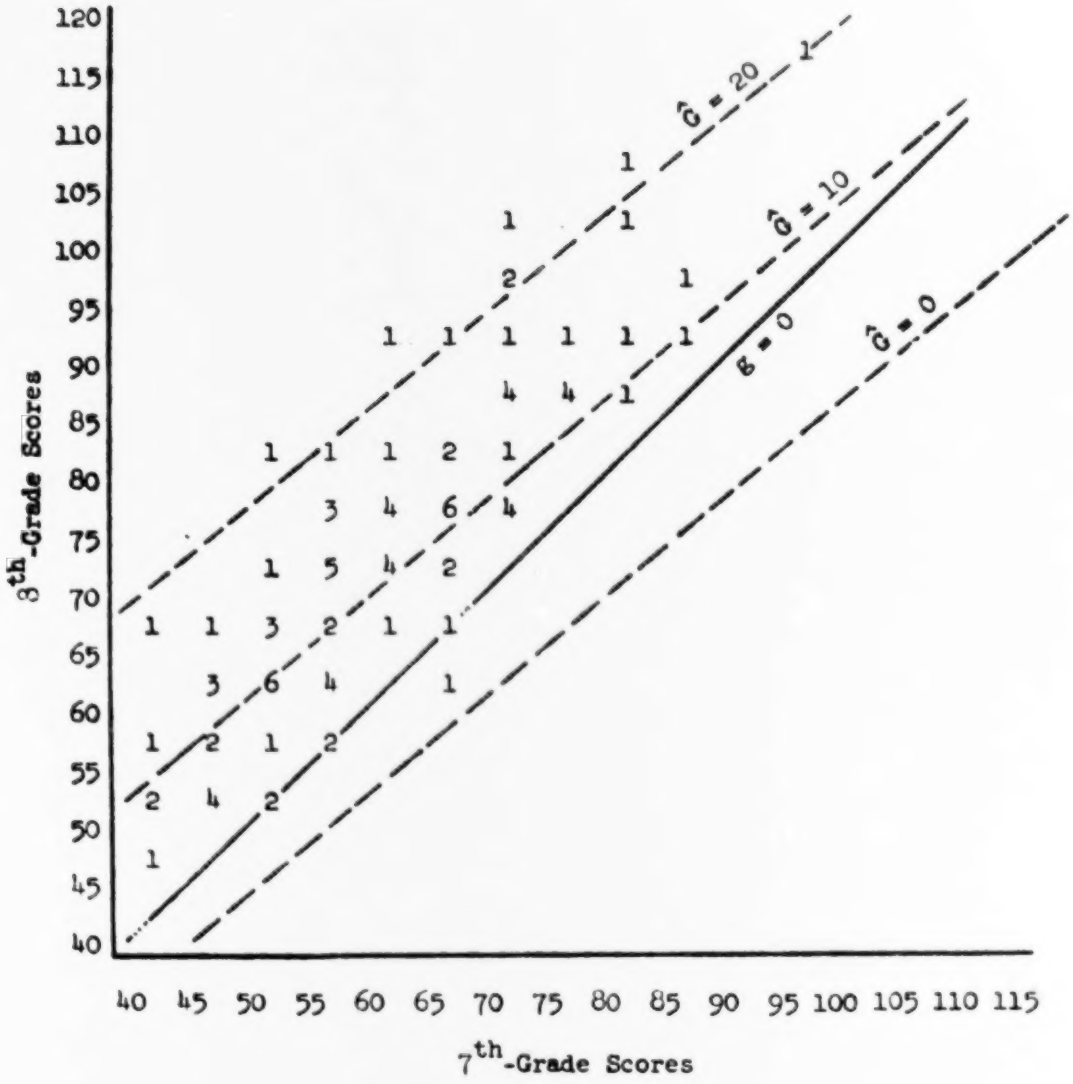


Figure 2

Scatterplot Showing Growth of 95 Students in One Junior High School Class (Courtesy Dr. William E. Coffman)

EXHIBIT IV

(Olsen, S.A. Doctoral dissertation, University of Iowa, 1981, p.45)

Figure 3: School A's Progress Points
Within the District Reference Group
(Circled points are those of School A)

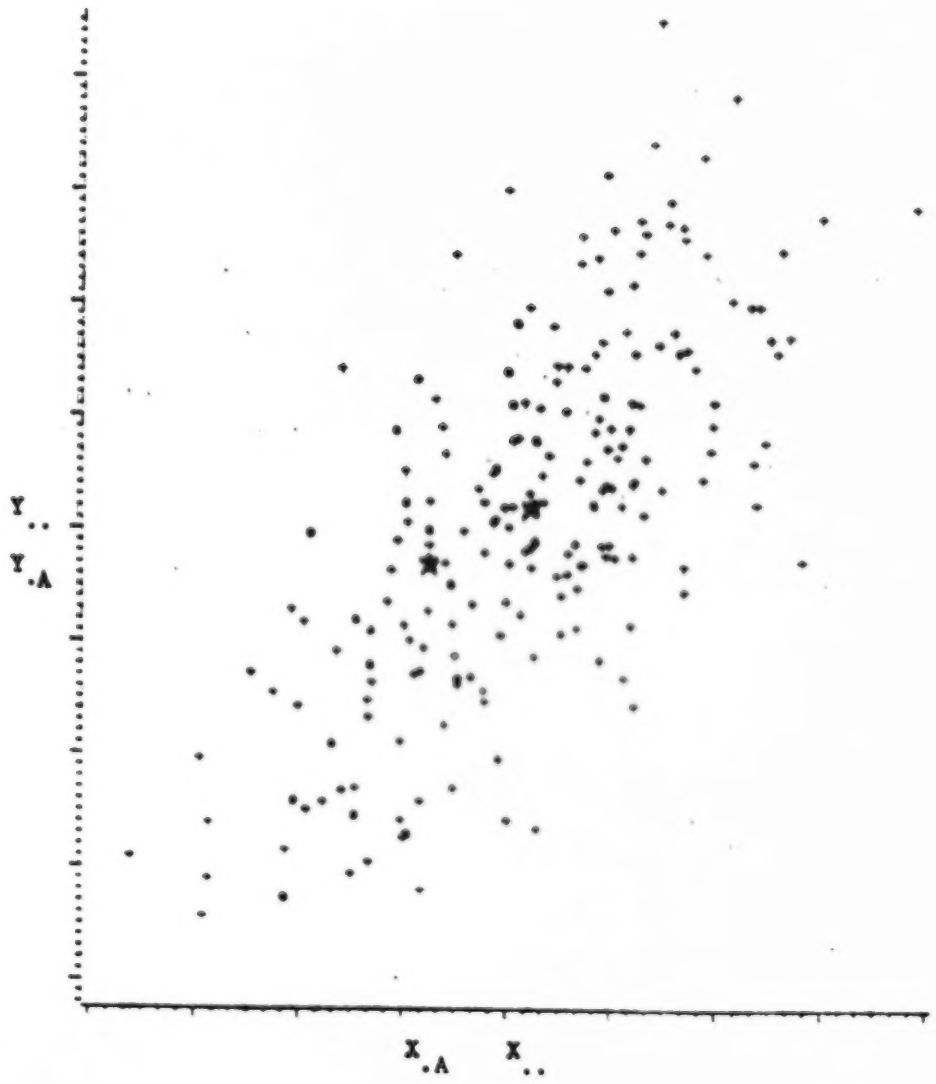
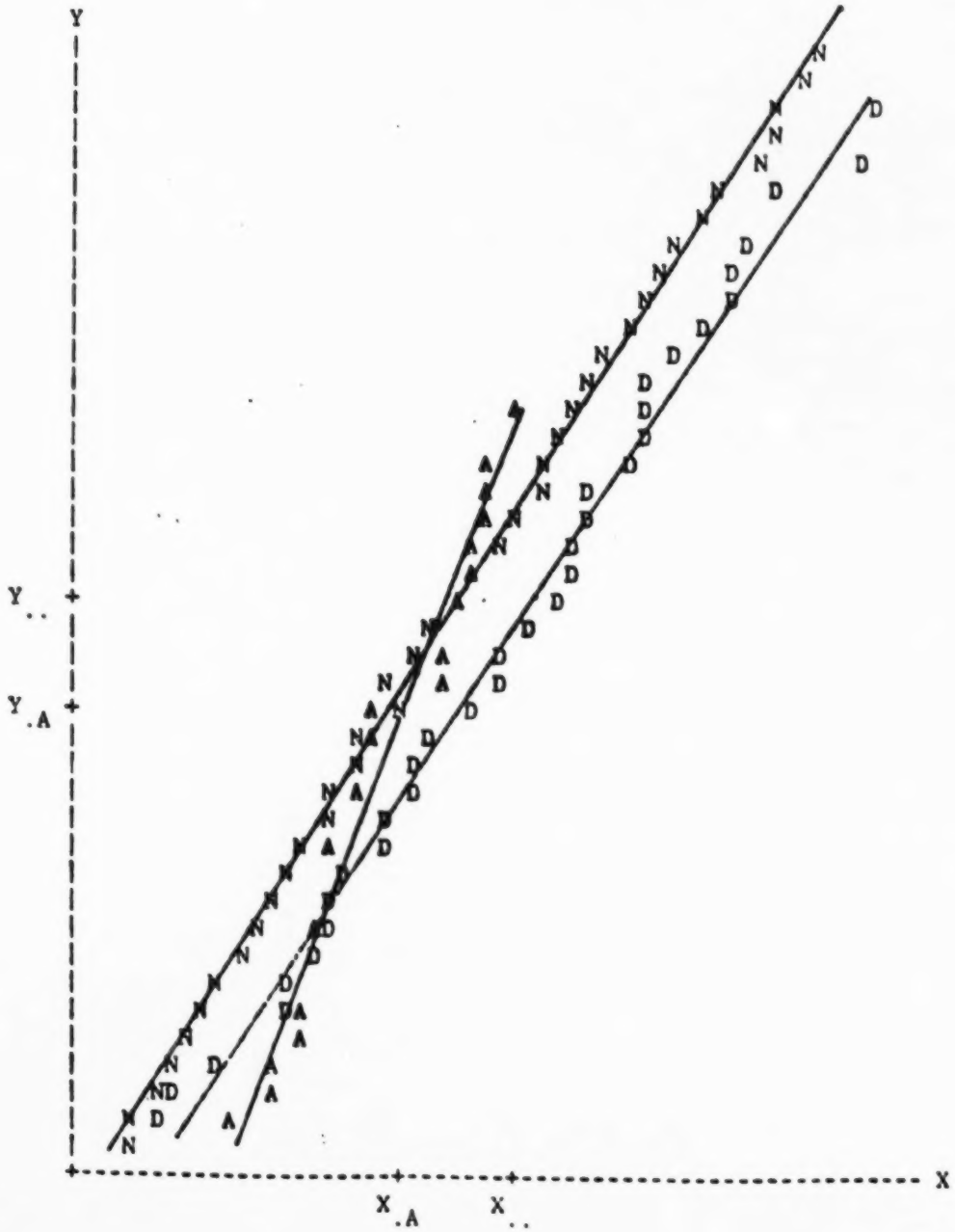


EXHIBIT V

(Olsen, S.A. Doctoral dissertation, University of Iowa, 1981, p.47)

Figure 4: Points of Equal Rank For
School A, District D and National Norm N



REFERENCES

Chen, Lin-Shua (1974). A study of the stability of patterns of academic growth for two samples of schools. Unpublished master's thesis, University of Iowa, Iowa City, IA.

Coffman, W.E. (1960). Patterns of growth in basic skills in two elementary school classrooms over a four-year period. Seventeenth Yearbook, National Council on Measurements Used in Education, pp. 141-151.

Coffman, W.E. (1987). Measurement of thinking skills - An historical perspective. Unpublished manuscript.

Cook, W.W. (1951). The function of measurement in the facilitation of learning. In E.F. Lindquist (Ed.), Educational Measurement (pp. 3-46). Washington, DC: American Council on Education.

Cronbach, L.J. & Furby, L. (1970). How should we measure "change" -- or should we? Psychological Bulletin, 74, 68-80.

Dyer, H.S. & Patton, M.J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 6, 591-605.

Fish, L.J. (1930). Examinations seventy-five years ago and today: Comparison and results of entrance examinations to high school. Yonkers-on-Hudson, NY: World Book Co.

Frederick, J.M. (1987, January). Measuring school effectiveness: Guidelines for educational practitioners (ERIC/TME) Report 93). Princeton, NJ: ERIC Clearinghouse on Tests, Measurements, and Evaluation, Educational Testing Service.

Harris, C.W. (Ed.). (1963). Problems in measuring change. Madison, WI: University of Wisconsin Press.

Jaeger, R.M. (1987). Two decades of revolution in educational measurement? Educational Measurement: Issues and Practices, 6(4), 6-14.

Ketcham, W.A. & Lafitte, R.G. (1959). How well are they learning? Educational Leadership, 16, 337-341, 350.

Kirst, M.W. & Wirt, F.M. (1982). State influences on education, In H.E. Mitchel (Ed.). Encyclopedia of educational research, Fifth edition (pp. 1770-1780). New York: The Free Press.

Learned, W.S. & Wood, B.D. (1938). The student and his knowledge. New York: The Carnegie Foundation for the Advancement of Teaching, Bulletin 29.

(Coffman, 1988)

Lindquist, E.F. (1951). Preliminary considerations in objective test construction. In E.F. Lindquist (Ed.), Educational measurement, (pp. 119-158). Washington, DC: American Council on Education.

Lord, F.M. (1956). The measurement of growth. Educational and Psychological Measurement, 16, 421-437.

Lord, F.M. (1967). A paradox in the interpretation of group performance. Psychological Bulletin, 68, 304-305.

McNemar, Q. (1958). On growth measurement. Educational and Psychological Measurement, 18, 47-55.

Malin, J.E. (1930, Jan.-Mar.). History of the measurement movement. Educational Outlet, 4, 72-80, 149-155.

Monroe, W.S. (Ed.). (1950). Encyclopedia of educational research (rev. ed.). New York: Macmillan.

Olsen, S.A. (1982). The diagonal approach: an easily applicable method for evaluating the progress of intact groups using standardized test results. (Doctoral dissertation, University of Iowa, 1981). Dissertation Abstracts International, 42(11), 4801A. (University Microfilms No. DA8210023)

Pace, C.R. (1972). Thoughts on evaluation in higher education. Iowa City, IA: American College Testing Programs.

Peterson, P.L. (1982). Individual differences. In H.E. Mitzell (Ed.), Encyclopedia of educational research, Fifth edition (pp. 849-851). New York: The Free Press.

Scates, D.E. (1941). Fifty years of objective measurement and research in education. Journal of Educational Research, 41, 241-64.

Tyler, R.W. (1932). Service studies in higher education. Bureau of Educational Research Monographs. No. 15. Columbus: Ohio State University.

Werts, C.E. & Linn, R.L. (1969). Lord's paradox: A generic problem. Psychological Bulletin, 72, 423-425.

Werts, C.E. & Linn, R.L. (1971). Analyzing school effects: ANCOVA with a
lible covariate. Educational and Psychological Measurement, 31, 95-104.

Woods, B.D. (Ed.). (1933). Educational tests and their uses. Review of Educational Research, 3(1).